IASSIST \ DCN
Data Curation Workshop

IASSIST 2018 Annual Conference
30 May 2018
Presenters: Jennifer Moore, Mara Blake, Mara Sedlins, Wendy Kozlowski

DCN Overview

Rise of the Data Sharing Culture

Researchers are increasingly required/incentivised to share data

- Funder data sharing mandates
- Journal data sharing policies
- Disciplinary practices → emphasis on transparency and reproducibility

Data sharing is not a straightforward process!

Goal of data curation ⇒ Prepare and maintain research data in ways that make it useful beyond its original purpose, ensure completeness, and facilitate long-term reuse and citability.

Data curation = metadata, documentation, access, preservation, and more...

Well-curated data are...

- Easier for fellow scholars and future collaborators to understand
- More likely to be trusted
- The research they represent are more likely to be reproducible
- More likely to be properly cited
- Represent potential cost-savings
- Findable, accessible, interoperable, and reusable, or FAIR (Wilkinson et. al, 2016)

Role of libraries in data curation

Libraries and academic-based data repositories are just one piece of the data repository landscape.

TABLE 4.3
Examples of kinds of data repositories found in the United States.

Kind of Repository	Examples
Federally Funded Data Centers	NASA Distributed Active Archives (DAAC), NOAA National Centers for Environmental Information (NCEI), National Snow and Ice Data Center (NSIDC), USGS Earth Resources Observation Systems (EROS) Data Center (EDC)
Federally Funded Research and Development Centers (FFRDC)	National Center for Atmospheric Research (NCAR), Jet Propulsion Lab (JPL), Oak Ridge National Laboratory (ORNL)
National Libraries	National Library of Medicine (NLM), National Agricultural Library (NAL), Library of Congress (LOC)
State and Local Agencies	State geological surveys, County planning offices
Thematic Repository	Long Term Ecological Research Network Information System (LTER NIS), Andrews Forest LTER (AND), National Snow and Ice Data Center (NSIDC), Maria Rogers Oral History Program
Domain Repository	Global Biodiversity Information Facility (GBIF), Inter-university Consortium for Political and Social Research (ICPSR), DataOne, Interdisciplinary Earth Data Alliance (IEDA)
Institutional Repository	Purdue University Research Repository (PURR), Data Repository for the University of Minnesota (DRUM)
Replication Repository	Dryad Digital Repository, Pangaea Data Library
Software Repository	GitHub, SourceForge
Commercial Archives	DigitalGlobe, Aerial photography companies, Resource exploration companies, Figshare
Private Archives	Huntington Library, Getty Research Institute

Baker, K. and Duerr, R. (2017). “Data and a diversity of repositories” in *Curating Research Data: A handbook of current practice* (L. R. Johnston, ed.). ACRL press.

What is data curation?

Data curation is the active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education; curation activities enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time. (UIUC, 2007)

- Based in archival best practice (libraries know how to do this!)
- Data repositories provide a technological foundation
- But many curation activities are not easily automated ⇒ need curators (people)

Current state of libraries and data curation

- Surveyed 124 Association of Research Libraries (ARL) institutions in January 2017
- 80 institutions (65%) responded
- Goal: Understand the current data curation services offered and level of demand.

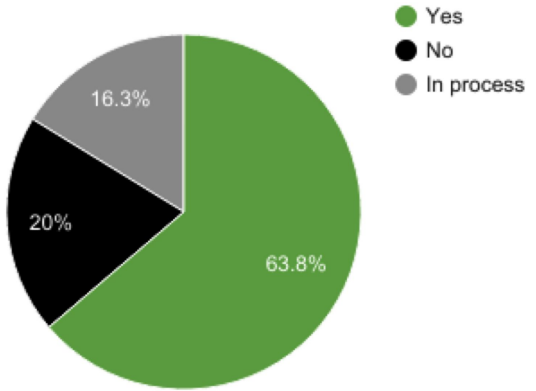
Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

Result: Nearly two thirds (51/80) provided data curation services, 13 planning services, 16 not

Of those that provided data curation services:

- **Recent/new service**: More than half began in 2010 or later.
- **Repository-focused**: Nearly all also provided repository services
- **Based in IR**: More than half had an institutional repository that accepted data and a few had a stand-alone data repository.
- **Platforms** ranged from: DSpace (22), Fedora/Hydra (10), Islandora (7), Custom solution (7), Dataverse (local installation) (7), Digital Commons/BePress (5), Dataverse (hosted) (4), Other platform (10) such as HUBzero, Open Science Framework, Rosetta, and SobekCM.

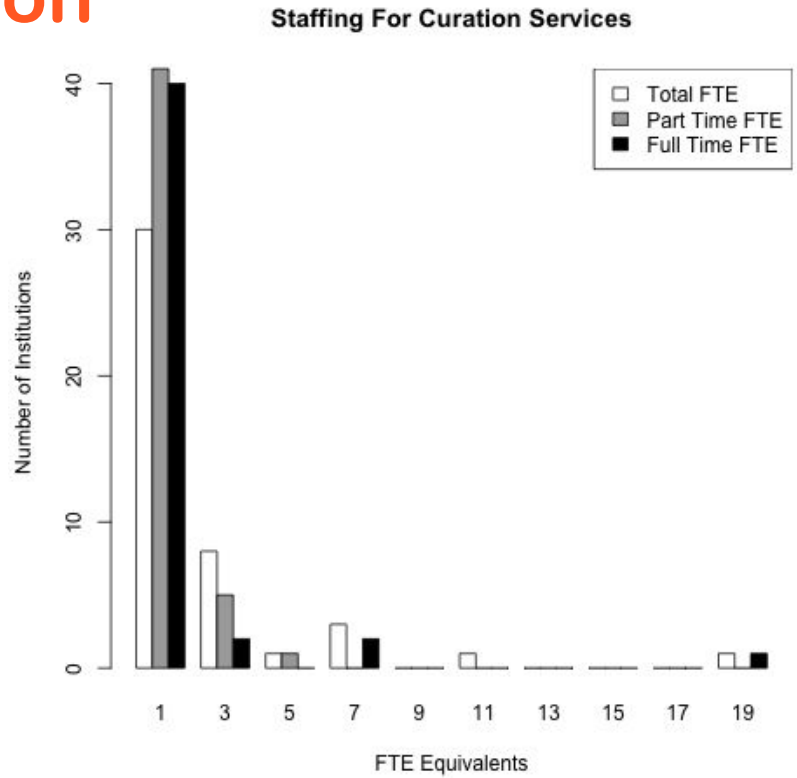
Does your institution currently provide research data curation services?



Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

Staffing for Data Curation

- Total Full-time equivalent (FTE) averaged to one person per institution dedicated to data curation services.
- Most libraries had 1 or more individuals providing data curation services at 5%-50% of their time while also carrying out other duties (part time FTE).



Cynthia Hudson-Vitale, Heidi Imker, Lisa R. Johnston, Jake R. Carlson, Wendy Kozlowski, Robert K. Olendorf, and Claire Stewart. (2017). SPEC Kit #354: Data Curation. Association of Research Libraries (ARL). May 2017. http://publications.arl.org/Data-Curation-SPEC-Kit-354/~~FreeAttachments/Data-Curation-SPEC-Kit-354.pdf

Challenge for Institutional Data Curation Services

How to scale data curation services across all disciplines?

Multiple experts are needed to effectively curate the diverse data types an institution typically generates.

Data curation expertise needed:

- File format-- GIS, spreadsheet/tabular, statistical/survey, software code, video/audio, images/3D, simulations...
- Discipline-specific-- genomic sequence, chemical spectra, biological image...
- Frequency-- Centers of excellence, departmental concentration

The Data Curation Network (DCN)
addresses this challenge by
collaboratively sharing data curation staff
across a network of
partner institutions and
data repositories.

Mission of the DCN

The Data Curation Network will enable academic institutions to better support researchers that are faced with a growing number of requirements to ethically share their research data.

In the next 3-5 years we will...

1. Develop standards-driven data curation techniques for all types of repository workflows and infrastructure.
2. Expand into a sustainable entity that grows beyond our initial six partner institutions.
3. Datasets curated by the Data Curation Network will be used to advance research and education in ways that are measurably of greater reuse value than non-curated data.
4. Build an innovative community that enriches capacities for data curation writ large.

Curation Training Pilot - Dec 2017, St. Louis, MO

Washington University in St. Louis

RESEARCH | DATA | GIS ABOUT US SERVICES RESOURCES PROJECTS COMMUNITY NEWS EVENTS

RESEARCH | DATA | GIS > DATA CURATION WORKSHOP

DATA CUR

DATA CURATION WORKSHOP

SLIDES AND HANDOUTS

DATE: DECEMBER 11 & 12, 2017

TWEET: #DCW2017

LOCATION: WASHINGTON UNIVERSITY IN ST. LOUIS, **MCMILLAN HALL,** ST. LOUIS, M

DESCRIPTION:

This free, 1.5 day workshop is open to all library staff and data professionals who are interested in data curation.

Participants will learn practical, hands-on treatments for data curation based on the **Data Curation Network CURATE** mod

- **C** – Check data files and read documentation;
- **U** – Understand the data (try to), if not...
- **R** – Request missing information or changes;
- **A** – Augment the submission with metadata for findability;
- **T** – Transform file formats for reuse and long-term preservation;
- **E** – Evaluate and rate the overall submission for FAIRness.

ATTENDEES WILL COME AWAY WITH:

1. A customized, implementable plan to enhance data curation activities at your local institution or organization,
2. Stakeholder focused talking points related to the value of data curation activities,
3. An in-depth understanding of specialized data curation practices in various disciplines, data types, and formats.

Learning Outcomes

1. A customized, implementable plan to enhance data curation activities at the local institution or organization,
2. Stakeholder focused talking points related to the value of data curation activities,
3. An in-depth understanding of specialized data curation practices in various disciplines, data types, and formats.

Curriculum Outline

1) Foundations of Curation overview
2) **CURATE** model training
3) Panel
4) Student project

Curate Model

- **C** – Check data files & read documentation
- **U** – Understand the data (try to), if not…
- **R** – Request missing information or changes
- **A** – Augment submissions with metadata \
- **T** – Transform for reuse & long-term preservation
- **E** – Evaluate & rate the submission for **FAIR**ness.

Check list

Checklist of C

CHECK Step

CURATE Action	Curator Checklist
Check data files and read documentation • Review the content of the data files (e.g., open and run the files or code). • Verify all metadata provided by the author and review the available documentation.	❑ Files open as expected ❑ Issues _____ ❑ Code runs as expected ❑ Produces minor errors ❑ Does not run and/or produces many errors ❑ Metadata quality is rich, accurate, and complete ❑ Metadata has issues _____ ❑ Documentation Type (circle) Readme / Codebook / Data Dictionary / Other: _____ ❑ Missing/None ❑ Needs work

UNDERSTAND Step

CURATE Action	Curator Checklist
Understand the data (or try to) • Check for quality assurance and usability issues such as missing data, ambiguous headings, code execution failures, and data presentation concerns. • Try to detect and extract any "hidden documentation" inherent to the data files that may facilitate reuse. • Determine if the documentation of the data is sufficient for a user with similar qualifications to the author's to understand and reuse the data. If not, recommend or create additional	*Varies based on file formats and subject domain. For example,...* Tabular Data Questions (Microsoft Excel) ❑ Organization of data well-structured ❑ Not rectangular ❑ Split tables into separate tabs ❑ Headers/codes clearly defined ❑ Define headers ❑ Clarify codes used _____ ❑ Clarify use of "blanks" ❑ Clarify units of measurement ❑ Quality control clearly defined ❑ Unclear quality control ❑ Update/add Methodology

REQUEST Step

CURATE Action	Curator Checklist
Request missing information or changes • Generate a list of questions for the data author to fix any errors or issues.	*Narrative describing the concerns, issues, and needed improvements to the data submission.*

University of Michigan sample email to researcher:

Dear [name of the person identified as the contact for the data set as stated in the DBD metadata],

Thank you for depositing your data set, [title of the data set] to the library's Deep Blue Data repository.

After we receive a data set, we review it to ensure that the data sets we host are as complete, accessible and understandable as possible. We have reviewed your data set and have the following recommendations for you:

- Recommendation #1
- Recommendation #2
- Recommendation #3

AUGMENT Step

CURATE Action	Curator Checklist
Augment the submission • Enhance metadata to best facilitate discoverability. • Create and apply metadata for the data record, including descriptive keywords. • When appropriate, structure and present metadata in domain-specific schemas to facilitate interoperability with other systems.	❑ Discoverability sufficient ❑ Recommend (circle one) full-text index / file rename / file reorder / file descriptions / zip files into one archive Other _____ ❑ Keywords Sufficient ❑ Suggestions _____ ❑ Linkages Sufficient ❑ Link to report/paper ❑ Link to related data sets ❑ Link to source data ❑ Link to other _____

TRANSFORM Step

CURATE Action	Curator Checklist
Transform file formats • Identify specialized file formats and their restrictions (e.g., Is the software freely available? Link to it or archive it alongside the data). • Transform files into open, non-proprietary file formats that broaden the potential audience for reuse and ensure that preservation actions might be taken by the repository in later steps. Retain original files if data transfer is not perfect.	❑ Preferred file formats in use ❑ Recommend conversion from _____ to _____ ❑ Retain original formats ❑ Software needed is readily available ❑ Unclear version of software ❑ Unclear software used ❑ Visualization of data easily accessible ❑ Recommend graphical representation _____ ❑ Recommend web-accessible surrogate _____

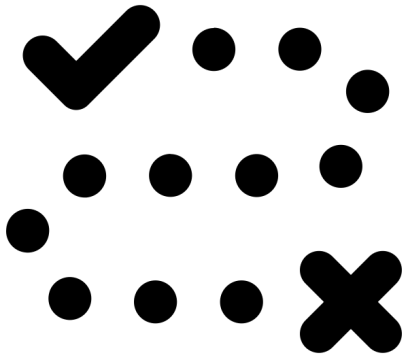
Cornell's List of Preservation Format Recommendations
http://guides.library.cornell.edu/ecommons/formats

EVALUATE Step

CURATE Action	Curator Checklist
Evaluate and rate the overall data record for FAIRness * • Score the dataset and recommend ways to increase the FAIRness of the data and become "DCN approved."	Findable - ❑ Metadata exceeds author/ title/ date, ❑ Unique PID (DOI, Handle, PURL, etc.). ❑ Discoverable via web search engines. Accessible - ❑ Retrievable via a standard protocol (e.g., HTTP). ❑ Free, open (e.g., download link). Interoperable - ❑ Metadata formatted in a standard schema (e.g., Dublin Core). ❑ Metadata provided in machine-readable format (OAI feed). Reusable - ❑ Data include sufficient metadata about the data characteristics to reuse ❑ Contact info displayed if the direct assistance of the author needed. ❑ Clear indicators of who created, owns, and stewards the data. ❑ Data are released with clear data usage terms (e.g., a CC License).

* Rubric evaluating the FAIR principles are based on the scoring matrix by Dunning, de Smaele, & Böhmer (2017).

Curriculum - Foundations of Curation



Created by Gregor Cresnar
from Noun Project

Training Goals:

- Define data curation
- Describe why it's needed
- Explore library Involvement - **Spec Kit**
- Demonstrate where participants are now (pre-assessment)
- Identify where to place data curation efforts

Curriculum - Foundations

Created by Gan Khoon Lay
from Noun Project

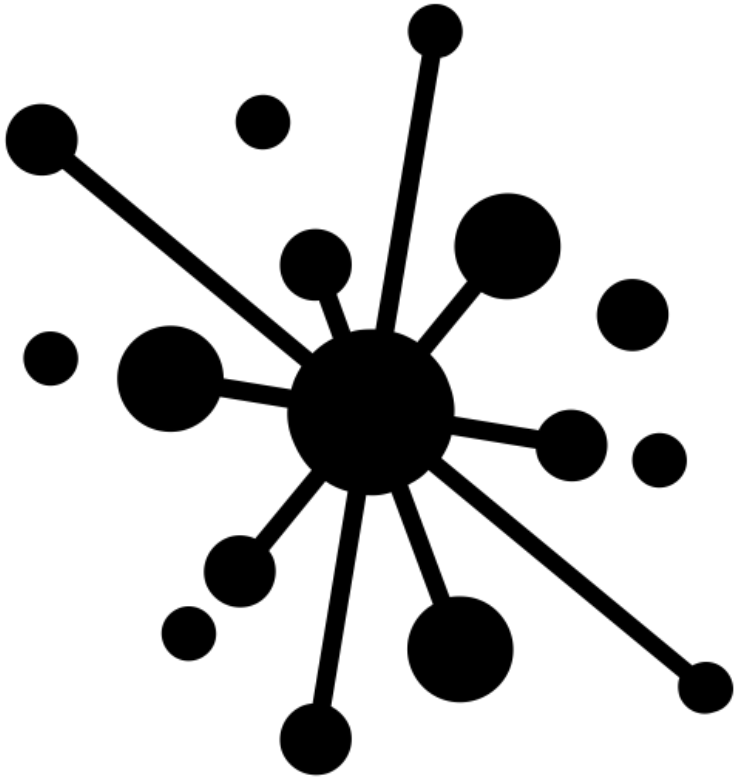
Explaining & Promoting Data Curation:

- Shared language for data curation
- Methodology for creating list of activities
- Importance of data duration activities
- Assessment & comparison
- Case for investment
- Gaps as opportunities for collaboration

Exercise: Developed elevator speeches to audience of choice

Curriculum - Datasets

1. Nematodes
2. Peer preferences
3. Chempolymers
4. Robotics & Locomotion
5. LGBTQ STL Locations



Created by Eliricon
from Noun Project

Curriculum - CU

C – **Check** data files & read documentation
U – **Understand** the data (try to), if not…

Created by Thomas' designs from Noun Project

Check

- files open
- code runs,
- metadata & documentation

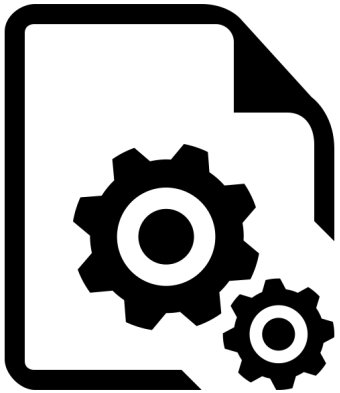
Understand

- quality assurance & usability issues
- hidden documentation
- documented for reuse

Exercise: Use checklist to check & understand example datasets

Curriculum - RA

R – **Request** missing information or changes
A – **Augment** submissions with metadata

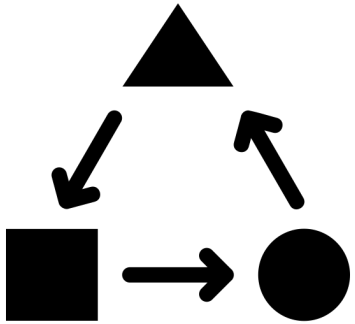


Created by Ilsur Aptukov
from Noun Project

- How to make requests to researchers (eg. simple, specific)
- Approaches to enhance or create metadata for findabiity
- Where possible, domain specific metadata

Exercise: 1) wrote a letter to submitter requesting info
2) reviewed dataset's metadata

Curriculum - T



Created by anbileru adaleru
from the Noun Project

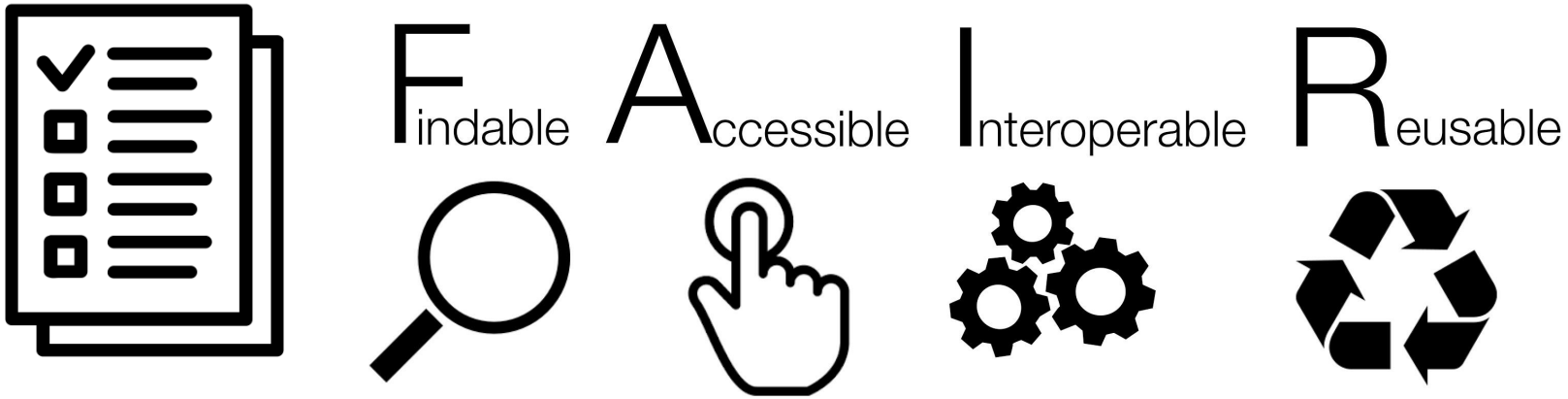
T – **Transform** for reuse & long-term preservation

- Concept of transformation
- Why transform?
- CS: GIS example (interoperability)
- CS: Excel example (preservation)
- CS: Video Transcription (discoverability)

Exercise: Brainstormed how to transform & transformation challenges

Curriculum - E

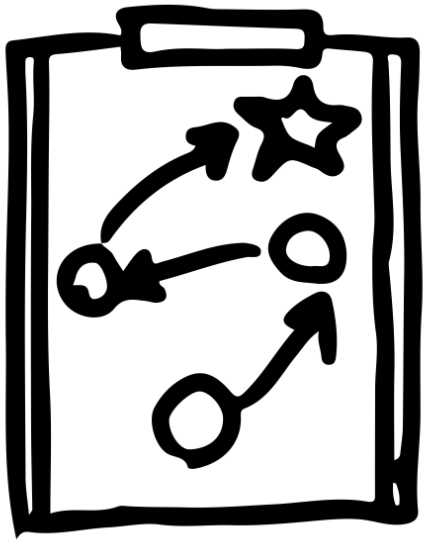
E – **Evaluate** & rate the submission for **FAIRness**.



Created by Yu Luck
from Noun Project

Exercise: Reviewed dataset for FAIRness & recorded what was missing

Curriculum - Student Project



Created by Valeriy
from Noun Project

- Reflected on the curation activities learned and applied in the workshop
- Each participant reflected on own organizational structure and institutional needs to identify:
  - 2 things to do at own institution
  - Expected challenge
  - Resources needed
  - Short-term goals

LESSONS LEARNED

REGARDLESS OF A PROJECT'S OUTCOME, WE CAN LEARN SOMETHING AND APPLY THAT LESSON TO THE NEXT PROJECT.

"I DIDN'T FAIL. I JUST FOUND A THOUSAND WAYS THAT DIDN'T WORK." - THOMAS EDISON

A few thoughts from Table 1

https://doi.org/10.13012/B2IDB-6946735_V2

Minute Paper

Students asked to anonymously respond to the following questions at the end of day one:

1. What was the most valuable thing you learned today?
2. What do you still have questions about?
3. Is there anything we could do to improve the workshop?

What do you still have questions about?

Category	Count
Standards/best practices	4
Humanities / Digital Preservation Content Curation	2
Interacting with researchers	3
Staffing	3
Software and tools	2
Total Responses	**27**

Content:

- More pre-workshop materials
- More hands-on work
- More time with software
- More on scientist perspectives

Physical/Logistics:

- More outlets, space, breaks
- Better acoustics
- Have presenters come in-person/ don't do remote presentations after lunch

Is there anything we could do to improve the workshop?

Category	Count
Pre-workshop content	2
Content	8
Physical/logistics	10
No suggestions	10
Total Responses	**30**

Content:

- More pre-workshop materials
- More hands-on work
- More time with software
- More on scientist perspectives

Physical/Logistics:

- More outlets, space, breaks
- Better acoustics
- Have presenters come in-person/ don't do remote presentations after lunch

Moving forward - IMLS

December workshop http://gis.wustl.edu/dgs/iassist-data-curation-workshop/

Laura Bush 21st Century Librarians

“Building the Digital Curation Workforce: Advancing Specialized Data Curation”

Goals:

1. expand functional data curation capabilities for librarians;
2. enhance the quality of data curated at institutions that participate in these programs; and
3. create discipline-specific or functional primers.

Schedule

Host three two-day workshops over the two year grant period:

1. October 17-18, 2018 in Las Vegas, NV, directly following DLF
2. Spring 2019 on the East Coast
3. Fall/Winter 2019 in the Midwest

More details coming soon!

Long-term Impact

The grant project will:

- Create intermediate to advanced data curation training modules and functional primers which will be made widely available for use and reuse.
- Enhance capacity for librarians nationwide to robustly curate the heterogeneous datasets created on their campuses.
- Expand the network of data curators, subjects specialists and repository managers nationwide.

Questions and Discussion